Significance Testing to Establish Equivalence Between Two Treatments in Matched Pairs Design

Jun-mo Nam
Biostatistics Branch, National Cancer Institute,
6120 Executive Blvd, EPS/8028, Rockville, MD 20892, U.S.A.

Abstract

When researchers want to establish that a new treatment is as effective as a standard one, the conventional test procedure is inapplicable. A null-hypothesis appropriate for this case is a specified non-zero difference between the response rates for the two treatments. The aim of such a study is to show that the difference between treatment effects is less than a prescribed small amount which is materially insignificant. The statistical test concerned with the equivalence of treatments is called an equivalence test.

For a matched pairs trial, the standard McNemar's test is inappropriate for establishing equivalence. Recently, the equivalence test for a one-sided question (e.g, non-inferiority) has been studied by several authors, e.g., Lu and Bean (1995, Statistics in Medicine 14, 1831-1839), Nam (1997, Biometrics 53, 1422-1430) and Tango (1998, Statistics in Medicine 17, 891-908). However, the test involving two-sided question (e.g., bioequivalence) has not been thoroughly investigated for data with binary outcome.

In this paper, we provide equivalence test for the two sided question by combining rejection regions of two one-sided tests. This is different from the conventional two-sided test. Likelihood score and Wald-type tests under this procedure are presented and the tests are numerically illustrated using actual data from a diagnostic study. A preference of the score method is indicated.

1. Introduction

The purpose of a conventional comparative study is to show that two treatments are not identical (e.g., a new treatment is more effective then the standard one) by rejecting a null-hypothesis of no difference in response rates. However, non-rejection of the null does not imply the confirmation of the null. The null hypothesis is never proved but it is possibly disproved in the course of experimentation (Fisher, 1935). The conventional formulation of testing hypothesis is inapplicable for the situation in which the aim of the trial is to establish equivalence of two treatments and demonstrate that the two treatments do not differ by more than a specified small amount. The proper null hypothesis is a specified non-zero difference between two treatments. This test procedure has been commonly called an equivalence test. Statistical methods for testing equivalence involving two independent binomial varieties in prospective studies have been investigated by many authors, e.g., Dunnet and Gent (1977), Blackwelder (1982), Rodary et al (1989), Farrington and Manning (1990), Yanagawa et al (1994) and Nam (1994, 1995). For matched-pairs studies, the McNemar test statistics (1947) has been used for detecting a difference between two treatments. However, the test is inappropriate for demonstrating equivalence. Recently, Lu and Bean (1995), Nam (1997) and Tango(1998) have investigated equivalence test and sample size requirement for matched-size pairs.

Equivalence tests in prospective studies, e.g., clinical trials are usually one-sided. The one-sided equivalence test procedure has been well established. However, an equivalence test for the two-sided question has not been thoroughly addressed. In a diagnostic study, for example, researchers are interested in whether an inexpensive and quicker scan method is as accurate as an expensive and time-consuming pathological examination in detecting lung disease. In this note, an appropriate two-sided equivalence test procedure which differs from the conventional two-sided one is formulated. We present Wald-type and likelihood score tests for two-sided equivalence for matched-pairs studies.

2. Testing Significance for Equivalence

Consider n pairs of matched samples (Y_{1j}, Y_{0j}) for j=1, 2, ..., n. The Y_{1j} and Y_{0j} are binary responses (e.g., + or !) of the new and standard treatments for the j^{th} pair. Four possible types of outcome of a matched-pair are (+, +), (+, !), (!, +) and (!, !). The observations and underlying probabilities are denoted by the following table

Observations				Probabilities			
	Standard				Standard		
New treatment	+	!	Sum	New treatment	+	!	Sum
+	x_{11}	x_{10}	$x_{1.}$	+	p_{11}	p_{10}	p_1
!	x_{01}	x_{00}	$x_{0.}$	!	p_{01}	p_{00}	q_1
Sum	$x_{.1}$	$x_{.0}$	n	Sum	p_0	q_0	1

Define the difference between new and standard treatments in response probabilities as $* = p_1 - p_0 = p_{10} - p_{01}$. Reparametrerize the multinomial probabilities as $p_{10} = p_{01} + *$ and $p_{11} + p_{00} = 1 - 2p_{01} - *$ and express the likelihood as

$$L = x_{10} \ln (p_{10} + *) + x_{01} \ln (p_{01}) + (x_{00} + x_{11}) \ln (1 - 2p_{01} - *) \quad (1)$$

where * is the parameter of interest and p_{01} is a nuisance parameter. Denote $\hat{p}_{ij} = x_{ij} / n$ for i, j = 0, 1, $\hat{p}_1 = x_{1.} / n$ and $\hat{p}_0 = x_{.1} / n$. Letting $t(*) = \hat{p}_{10} - \hat{p}_{01} - *$, we have the variance of t (*) as var $\{t (*)\} = (p_{10} + p_{01} - *^2) / n$. The maximum likelihood estimator (MLE) of p_{01} for a given value of *, $\tilde{p}_{01}$, is a solution of a partial derivative of (1) with respect to p_{01} for a given value of *, i.e., $(\partial L / \partial p_{01}) = 0$ which is a root of the quadratic equation:

$$\tilde{p}_{01} = \{-b + (b^2 - 4ac)^{1/2}\} (2a) \quad (2)$$

where $a = 2n$, $b = (2n + x_{01} - x_{10}) * - (x_{10} + x_{01})$ and $c = x_{01} * (1 - *)$.

2.1 One-sided Equivalence.

Consider an equivalence test for the one-sided question (e.g., non-inferiority test) and set the null and the alternative hypotheses as $* \# -*_0$ and $* > -*_0$ where $*_0$ is a specified small positive number. A Wald-type statistic for testing $* \# -*_0$ against $* > -*_0$ is obtained by a nominal deviate of $t (-*_0)$ calculated at $p_{10} = \hat{p}_{10}$ and $p_{01} = \hat{p}_{01}$, i.e.,

$$z_w (-*_0) = (x_{10} - x_{01} + n *_0) / (x_{10} + x_{01} - n *_0^2)^{1/2} \quad (3)$$

(e.g., Lu and Bean, 1995). We reject H_0^- at $\forall$ when $z_w(!*_0) \exists z_{(\forall)}$ where $z_{(\forall)}$ is the 100% x (1 !$\forall$) percentile point of the standardized normal distribution. The above test is anti-conservative, i.e., its false positive error rate is greater than a nominal $\forall$. Alternatively, using Bartlett's principle (1953), Nam (1997) derived the likelihood score test statistic which can be simplified as

$$z_w(!*_0, \hat{p}_{01}) = (x_{10} !x_{01} + n *_0)/ \{n(\tilde{p}_{10} + \tilde{p}_{01} !*_0^2)\}^{½} \quad (4)$$

where $\tilde{p}_{01}$ is the MLE of p_{01} for a given value of $*_0$ and $\tilde{p}_{10} = \tilde{p}_{01} !*_0$. The MLE, $\tilde{p}_{01}$, is found from (2) with $* = !*_0$. We reject H_0^- in a favor of H_1^- when $z_s(!*_0, \tilde{p}_{01}) \exists z_{(\forall)}$. The score test possesses asymptotically optimum properties. The type 1 error probability of the test is satisfactorily close to the nominal level for a small sample size. McNemar's statistic is a special case of the statistics, (3) and (4), for $*_0 = 0$.

2.2 Two-Sided Equivalence.

For two-sided equivalence (e.g., bioequivalence), we may formulate the null and alternative as

H_0: *** $\exists *_0$ against H_1: *** $< *_0$,

which is the same to the following two one-sided hypotheses:

H_0^-: *# $!*_0$ against H_1^-: *> $!*_0$,

H_0^+: *$\exists *_0$ against H_1^+: *$< *_0$.

Note that two null hypotheses are disjointed intervals. We reject H_0 if and only if both of H_0^- and H_0^+ are rejected. Define rejection regions of the two one-sided tests involving H_0^- vs H_1^- and H_0^+ vs H_1^+ as R^- and R^+, respectively. Then, the overall rejection region for H_0 is the intersection of the regions, i.e., R^- 1 R^+. Since the probability of the intersection of the two rejection regions is less than or equal to the probability of either rejection region, the test with the rejection region, R^- 1 R^+, has at most $\forall$ if each of the two one-sided tests associated with R^- and R^+ has $\forall$ (Berger, 1982, and Hsu, et al, 1994). Consider a one-sided test for H_0^+ against H_1^+. Wald-type and score statistics are

$$z_w(*_0) = (x_{10} !x_{01} ! n *_0)/(x_{10} + x_{01} !n *_0^2)^{½} \quad (5)$$

$$z_s(*_0, \tilde{p}'_{01}) = (x_{10} - x_{01} - n*_0) / \{n(\tilde{p}'_{10} + \tilde{p}'_{01} - \delta_0^2)\}^{½} \quad (6)$$

where $\tilde{p}'_{01} = \{!b\rho + (b\rho^2 !4ac\rho)^{½}\}/(2a)$ with $b\rho = (2n + x_{01} - x_{10})*_0 !(x_{10} + x_{01})$ and $c\rho = x_{01} *_0 (1 !*_0)$ and $\tilde{p}'_{10} = \tilde{p}'_{01} + *_0$ from (2). The Wald test procedure for H_0 against H_1 is to reject H_0 at $\forall$ when $z_w(!*_0) \exists z_{(\forall)}$ and $z_w(*_0) \# !z_{(\forall)}$ from (3) and (5), or do not reject H_0 otherwise. Similarly, we may establish equivalence using the score method when $z_s(!*_0, \tilde{p}_{01}) \exists z_{(\forall)}$ and $z_s(*_0, \tilde{p}'_{01}) \# !z_{(a)}$ from 4) and (6). In this section, we consider two-sided equivalence as $*** < *_0$ where $*_0$ is a small positive number.

We may generalize test two one-sided testing hypotheses as

$H_0^- : * \# !*_0$ against $H_1^- : * > !*_0$,

$H_0^+ : * \# *'_0$ against $H_1^+ : * < *'_0$

where $*_0$ and $*'_0$ are positive and may differ. If we reject H'_0 at $\forall$ and also H_0^+ at $\forall'$, then the above two null hypotheses are rejected at most at a level of maximum of $\forall$ and $\forall N$. Analogously, we can obtain Wald and score test statistics for the generalized hypothesis.

3. An Example

A study on use of liver scans and pathological examination for detecting liver disease in 344 patients (McNeil, Keeler and Adelstein, 1975) is summarized in the following table:

	Pathological exam.		
Scan	+	-	Sum
+	231	32	263
-	27	54	81
Sum	258	86	344

Liver disease rates diagnosed by scans and pathological examinations were $\hat{p}_1 = (231 + 32)/344 = 0.765$ and $\hat{p}_0 = (231 + 27)/344 = 0.75$, respectively. The disease rate by scan was greater than the morphological one by 1.5%. We want to find out whether the liver scan diagnosis is equivalent to the pathological test in terms on identifying the disease rate.

If we consider the absolute difference between two positive response rates less then 0.05 (i.e., $|*| < 0.05$) as equivalence, then we may set H_0: $|*| \exists 0.05$ against H_1:$|*| < 0.05$ for testing equivalence. The two one-sided Wald test statistics are z_w^- (!0.05)=2.911 (p=0.002) and z_w^+ (0.05)=!1.600 (p=0.054) from (3) and (5), respectively. Although the value of z_w^- is in its rejection region, the z_w^+ isn't. Therefore, we cannot reject H_0 at $\forall$ =0.05 when the threshold value of equivalence is $*_0 = 0.05$. For the score method, we calculate the MLE of p_{01} for * = !0.05 as $\tilde{p}_{01} = 0.1179$ and that for * = 0.05 as $\tilde{p}_{01}^{'}$ =0.0646. The two one-sided score statistics are z_s^- (!0.05, 0.1179)= 2.796 (p=0.003) and z_s^+ (0.05, 0.0646) = !1.565 (p=0.06) from (4) and (6), respectively. The score method, also, does not indicate two-sided equivalence defined as above. The p-value of the Wald tests are smaller than those of the score tests as expected since the Wald method is anti-conservative.

If underestimation of disease rate is more serious than its overestimation, then we may set the practical equivalence as $(!*_0, *_0^{'})$ where $0 < *_0 < *_0^{'}$ (a non-equi-distance from zero), say, (-0.03, 0.06). One-sided Wald test statistics for H_0^- : *# !0.03 versus H_1^- : * > !0.03 and that for H_0^+ : * $\exists$ 0.06 versus H_1^+ : * < 0.06 are z_w^- (-0.03) = 2.000 (p=0.023) and z_w^+ (0.06) = -2.058 (p=0.020), respectively. Based on these two one-sided tests, we reject H_0: * $\notin$ (-0.03, 0.06) in a favor of H_1: *, (-0.03, 0.06). The MLE's of a nuisance parameter are $\tilde{p}_{01}$ = 0.0619 and $\tilde{p}_{01}^{'}$ =0.1838 for * = -0.03 and *=0.06. The two one-sided score statistics are z_s^- (-0.03, 0.0619) = 1.921 (p=0.029) and z_s^+ (0.06, 0.1838) = !1.986 (p=0.024).

Both score tests reject their corresponding null hypothesis at $\forall$= 0.05 and we may conclude that the two diagnostic methods are practically equivalent when the equivalence is defined as $(!*_0, *_0^{'})$=(-0.03, 0.06).

4. Remarks

The two-sided equivalence test procedure is different from the conventional test procedure. The former postulates the null hypothesis that the difference in response rates by two treatments is outside of or on a neighborhood of zero difference and the alternative that the difference is in the neighborhood. We can establish equivalence of the treatments by

rejecting the null in a favor of the alternative. Specific values related to a boundary of the neighborhood, e.g. $*_0$ and are $*'_0$, are very important in inference on equivalence and testing a significance. The value should be reasonable and acceptable for the specific subject matter, and in accordance with the purpose and the nature of a study. The overall rejection region of the two-sided equivalence test is the intersection of the rejection regions of two one-sided tests while the rejection region of the conventional two-tailed test is the union of rejection regions of two one-tailed tests. Therefore, the type 1 error probability of the two-sided equivalence test is at most $\forall$ when the level of each one-sided test is $\forall$, but the level of the conventional two-tailed test is $2\forall$ when each one-tailed test is level $\forall$.

The Wald-type test is anti-conservative and its p-value is smaller than the nominal one. Actual type 1 error probability of the score test is satisfactorily close to the nominal level. We recommend the score test, particularly, for a small sample size. We can derive proper confidence intervals consistent with two-sided Wald and score tests for matched-pairs samples. It should be noted that such a confidence interval is quite different from a conventional interval. Two conventional one-tailed tests of size $\forall$ are associated with a $1-2\forall$ confidence interval. We can construct a $1-\forall$ confidence interval from two $\forall$-level one-sided equivalence tests and vice-versa. A paper dealing with this subject for matched samples is in preparation.

REFERENCES

Bartlett, M.S. (1953). Approximate confidence intervals. II: More than one unknown parameter. *Biometrika* **40**, 306-317.

Berger, R.L. (1982). Multiparameter hypothesis testing and acceptance sample. *Technometrics* **24**, 295-300.

Blackwelder, W.C. (1982). Proving the null hypothesis in clinical trials. *Controlled Clinical Trials* **3**, 345-359.

Dunnet, C.W. and Gent, M. (1977). Significance testing to establish equivalence between treatment, with a specific reference to data in the form of 2x2 tables. *Biometrika* **33**, 593-602.

Farrington, C.P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* **9**, 1447-1454.

Fisher, R.A. (1935). *The Design of Experiments. New York: Hafner.*

Hsu, J., Hwang, J.T.G., Lin, H. and Ruberg, S. (1994). Confidence interval associated with tests for bioequivalence. *Biometrika* **81**, 103-114.

Lu, Y. and Bean, J.A. (1995). On the sample size for one-sided equivalence of sensitivities based upon McNeman's test. *Statistics in Medicine* **14**, 1831-1839.

McNeil, B.J., Keeler, E. and Adlestein, S.J.(1975). Primer on certain elements of medical decision making. *New England Journal of Medicine* **239**, 211-215.

McNemar, S. (1947). Note on the sampling error of the differences between corrected proportions or percentages. *Psychometrika* **12**, 153-157.

Nam, J. (1994). Sampling size requirements for stratified prospective studied with null hypothesis of non-unity relative risk using the score test. *Statistics in Medicine* **13**, 79-86.

Nam, J. (1995). Sample size determination in stratified trials to establish the equivalence of two treatments. *Statistics in Medicine* **30**, 2037-2050.

Nam, J. (1997). Establishing equivalence of two treatments and sample size requirements in matched-pairs design. *Biometrics* **53**, 1422-1430.

Rodary, C., Com-Nougue, C. and Tourade, M. (1989). How to establish equivalence between treatments: A one-sided clinical trial in pediatric oncology. *Statistics in Medicine* **8**, 593-598.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample-design. *Statistics in Medicine* **17**, 891-908.

Yanagawa, T., Tango, T. and Heijima, Y. (1994). Mantel-Haeszel-type tests for testing equivalence or more than equivalence in comparative clinical trials. *Biometrics* **50**, 859-864.